Robust Anomaly Detection Under Contaminated Data: A Comprehensive Evaluation Across PHM Contexts

Stefano Donné[*†] Jesse Davis[†] Bart De Clerck[*] Mathias Verbeke[†‡]
[*] *MWMW, Department of Mathematics, Royal Military Academy (RMA), Belgium.*
[†] *DTAI, Department of Computer Science, KU Leuven, Belgium.*
[‡] *Flanders Make@KU Leuven, Belgium.*
stefano.donne@kuleuven.be, jesse.davis@kuleuven.be, bart.declerck@mil.be, mathias.verbeke@kuleuven.be

ABSTRACT

Robust anomaly detection under contaminated training data is an important challenge in Prognostics and Health Management (PHM). In semi-supervised anomaly detection, models are typically trained on data assumed to represent normal behavior. In practice, this "normal" set often contains an unknown fraction of abnormal or degraded samples, which can harm diagnostic performance. This work presents a comparative evaluation of several techniques designed to mitigate the effects of contaminated training data across four public datasets representative of diverse PHM contexts, spanning tabular and multivariate time-series data, as well as both discrete anomalies and gradual degradation processes. The results show that contamination-mitigating techniques can improve anomaly detection performance over classical baselines when constructing a training set consisting solely of normal instances is not feasible. However, the benefits offered by contamination-mitigating approaches vary according to dataset characteristics. The largest gains are observed on the time-series datasets considered here, suggesting that refinement techniques may offer a clearer advantage over contamination-sensitive baselines in these settings. These gains, however, come at a substantially higher computational cost. The experiments also suggest that the effect of contamination depends not only on its ratio, but also on the structure and distribution of anomalies.

1. INTRODUCTION

The Fourth Industrial Revolution (Industry 4.0) is defined by the integration of connected sensing, cloud computing, and artificial intelligence into modern industrial operations. Within this context, Predictive Maintenance (PdM) has emerged as a key capability: by combining sensor measurements with historical operating data, PdM aims to anticipate equipment failures and schedule interventions before costly unplanned downtime occurs (Carvalho et al., 2022).

Prognostics and Health Management (PHM) provides a structured framework for deploying PdM in practice. PHM typically combines diagnostics to assess the current health state of a system, with prognostics which estimate how that state will evolve over time. A cornerstone of PHM diagnostics is Anomaly Detection (AD), which flags observations that deviate from expected normal behavior and may indicate incoming faults or degradation. The most suitable AD approach depends strongly on the application domain, including the data modality (e.g., tabular data, images, or time series), the dimensionality or complexity of the underlying signals and the nature of the anomaly itself (e.g., point-wise, contextual, collective) (Chandola, Banerjee, & Kumar, 2009).

Semi-supervised methods, in particular reconstruction-based or forecasting techniques, have proven to be particularly effective in identifying anomalies within high-dimensional and time-series datasets (Wagner et al., 2023; Schmidl, Wenig, & Papenbrock, 2022). *Semi-supervised* AD methods do not learn to separate normal from abnormal data directly. Instead, they learn a representation of normal operating behavior and turn deviations from this representation into an anomaly score. In the literature, this semi-supervised AD setting is also commonly described as one-class classification (OCC) or novelty detection (Chandola et al., 2009). A canonical approach is to train an autoencoder (AE) on data that is assumed to be normal. Then, anomalies can be identified based on the heuristic that normal samples should have smaller reconstruction errors than abnormal examples.

However, in realistic PHM settings, truly clean training data are often unavailable. We refer to this as *contamination*: the presence of anomalous samples within the training set that is assumed to represent normal operating conditions. In PHM, contamination can arise for several reasons, including scarce labeling, imperfect maintenance logs, undetected incoming faults, operating condition changes, or data-quality events such as sensor issues and communication dropouts (del Moral, Nowaczyk, & Pashami, 2022; Kermenov, Nabissi, Longhi, &

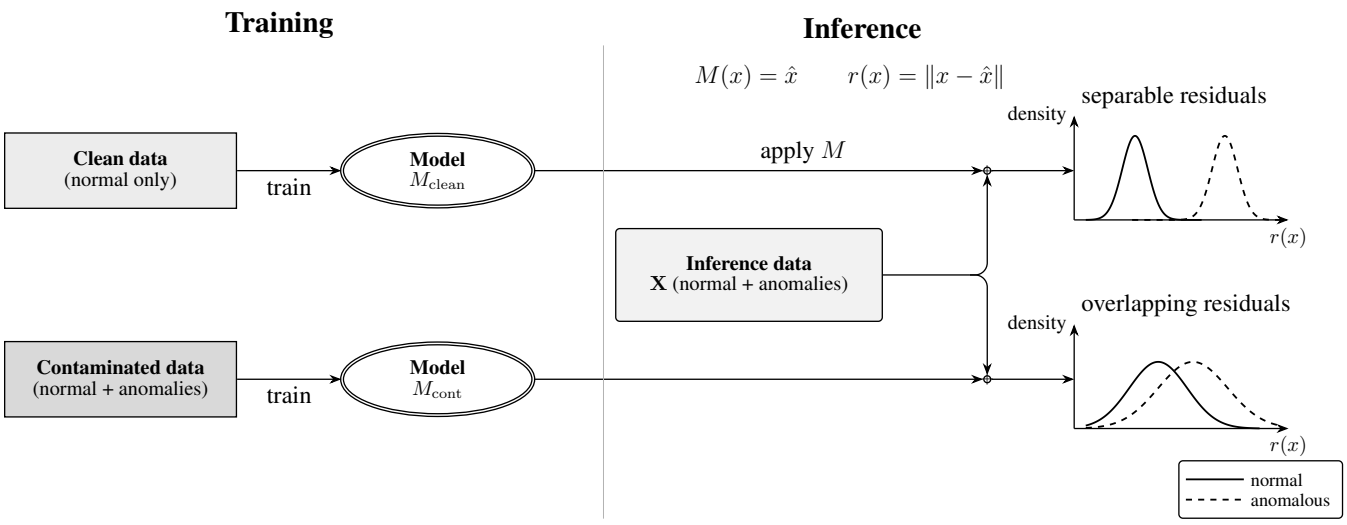

Figure 1. Impact of data contamination on reconstruction-based anomaly detection: a model trained on clean normal data yields separable residual distributions, while training on contaminated data can blur the normal/anomalous separation at inference.

Bonci, 2023; Huang, Tang, VanZwieten, & Liu, 2020; Zhao et al., 2019). When contaminated samples are used to fit a semi-supervised AD model (e.g., an AE), the model may inadvertently learn part of these abnormal patterns as normal, reducing the difference between normal and anomalous behavior at diagnosis time. For instance, (early) gradual degradation can be absorbed into the learned representation and soften the response of the model when encountering later degradation stages.

Ignoring contamination can degrade AD performance even at low contamination ratios (Munir, Siddiqui, Dengel, & Ahmed, 2019; Yu, Kim, Kim, & Oh, 2026; Qiu, Li, Kloft, Rudolph, & Mandt, 2022), although the magnitude of this degradation depends on the context. This is an issue in PHM because AD is often the first stage in a diagnostic pipeline, and errors at this stage can propagate to prognostics and downstream predictive maintenance decisions.

The present work benchmarks several anomaly-detection methods designed to mitigate contamination across different PHM settings with induced contamination. We consider five methods designed for contaminated settings, including three fully or partially model-agnostic frameworks implemented here with an autoencoder (AE) as the underlying one-class classifier. The AE is also evaluated as a standalone baseline, representing a semi-supervised method that is not designed to be robust to contamination. In addition, we include two classical anomaly-detection methods: one-class SVM (OCSVM) and Isolation Forest. We select four datasets that reflect PHM diversity, covering discrete and continuous anomaly phenomena in tabular and time-series data. In the first setting, methods are evaluated in a transductive outlier-detection setting, where model fitting and scoring are performed on the same unlabeled input pool, and performance is measured only on held-out labels. The second setting uses an inductive protocol, where each model is trained on a contaminated training $D_{-f,c}$ and evaluated on a separate held-out fold D_f . Our findings suggest that contamination-mitigating techniques can be beneficial in these settings; we further discuss several observations from our experiments and propose directions for future research.

2. RELATED WORK

Prior work on contaminated anomaly detection can be divided into PHM-oriented studies and broader contaminated-AD or OOD-detection studies. Table 1 summarizes their data setting, contamination assumptions, reproducibility, evaluation, and remaining gap. Overall, existing studies either focus on PHM with limited method or dataset breadth, or evaluate contamination-robust methods outside PHM, motivating our comparison across several mitigation families, public PHM datasets, and both discrete and gradual anomaly mechanisms.

3. METHODOLOGY

We designed two complementary comparisons to evaluate mitigation techniques for contamination across PHM contexts. Experiment 1 (Figure 2) is a transductive benchmark in which methods are fitted and scored on the same unlabeled input pool, while evaluation is performed only on held-out labels. Experiment 2 (Figure 3) is an inductive robustness study in which models are trained on a contaminated training split and evaluated on a fixed held-out fold as the contamination ratio increases.

Table 1. Positioning of prior work on anomaly detection with contaminated training data.

Work	Setting, anomalies, and contamination	Public data	Open code	Evaluation and gap
Marine OCC study (Tan et al., 2020)	PHM; tabular simulation data for marine propulsion degradation. Degradation-induced abnormal states are injected into the OCC training set; reported contamination ratios range from 0.05 to 0.50, plus a clean setting.	Yes	No	Reports condition-monitoring performance under contaminated training. The comparison is limited to classical OCC methods and does not evaluate contamination-mitigation mechanisms.
USDR (Ulmer et al., 2024)	PHM; industrial time-series data, including machine acoustics and turbofan degradation. Training data are unlabeled and may contain unknown abnormal samples; in one C-MAPSS setup, only 20% of normal samples are used, although contamination is not reported as a single anomaly fraction.	Yes	No	Reports refinement performance against blind contaminated training and a clean-data oracle. The method is relevant, but is not compared with other mitigation mechanisms and does not analyze anomaly-structure effects.
AD-DDPM (Wen et al., 2025)	PHM; rotating-machinery fault diagnosis on SQI and MGB fault-simulation datasets. Nominal monitoring data are contaminated by fault samples; reported contamination ratios are $\sigma \in \{0, 0.1, 0.2\}$.	No	No	Reports AUC, Acc, F1, TPR, and FAR, with high robustness under $\sigma = 0.2$. The diffusion-based method is not compared with other contamination-mitigation mechanisms.
WIRACAD (Donné et al., 2025)	PHM; degradation and health-monitoring benchmarks. Early or progressive degradation samples contaminate the data used to learn normal behavior; reported degraded/contaminated training proportions range from 63% to 76%.	Yes	Yes	Reports health-indicator quality with fit and monotonicity metrics. The study focuses on one iterative weighting mechanism; the present benchmark compares it with other robust methods across discrete and gradual anomaly settings.
SRR (Yoon et al., 2022)	Generic AD; image and tabular benchmarks. Fully unlabeled training sets contain both normal and anomalous samples; reported anomaly ratios include 10% on CIFAR-10 and 2.5% on Thyroid.	Yes	No	Reports AUC, average precision, and F1, with gains over one-class baselines. The method evaluates iterative sample removal, but not its behavior under gradual degradation or asset-wise time-series evaluation.
Iterative weighting (Kim et al., 2024)	Generic AD; tabular and image benchmarks. Contamination is modeled as anomalous samples in the training distribution and studied over reported ratios from 0.1% to 31.6%.	Yes	No	Reports AUROC across multiple generic AD and image benchmarks. The work supports iterative reweighting under contamination, but does not compare against other contamination-mitigating techniques.
LOE (Qiu et al., 2022)	Generic AD; image, tabular, and video benchmarks. Contaminated unlabeled training data are handled by inferring latent normal/anomalous labels; reported ratios include 10% on CIFAR-10/F-MNIST and tabular datasets, 10–20% on MVTec, and 5–20% sensitivity studies.	Yes	Yes	Reports AUC and F1 across several benchmarks and backbones. Its latent-label/outlier-exposure framing differs from the controlled one-class setting considered here, and discrete-vs-gradual robustness is not isolated.
RDA (Zhou & Paffenroth, 2017)	Generic AD; MNIST image data. Training data may contain outliers and sparse corruptions; the outlier-detection setup mixes 265 anomalous digits with 4859 nominal digits, yielding about 5.2% anomalies.	Yes	Yes	Reports denoising error and precision/recall/F1 for outlier detection; RDA reaches F1 = 0.64 versus 0.37 for Isolation Forest. It is a useful robust-AE baseline, but does not evaluate contamination sweeps across mitigation mechanisms.
READ (Shou et al., 2025)	Generic AD; multi-domain public datasets with contaminated unlabeled data and limited labels. The method selects informative subsets containing potential anomalies and difficult normal samples; a 10% contamination setting is reported.	Yes	Yes	Reports robustness and efficiency under contamination with limited supervision. The subset-selection mechanism is relevant, but the labeled-clean-data assumption motivates evaluating both informed and unsupervised variants.
OOD reweighting (Li et al., 2025)	OOD detection; image benchmarks. OOD samples contaminate the in-distribution training set; reported OOD-contamination settings include 1%, 2%, and 5%, with continuous reweighting used to reduce their influence.	Yes	No	Reports AUROC and OOD-percentage estimation error. The objective is OOD detection under contaminated in-distribution training, whereas this paper evaluates anomaly-score robustness under contaminated normal modeling.
NCAE (Yu et al., 2026)	Generic AD; image benchmark datasets such as MNIST and Fashion-MNIST. Training data are contaminated by samples from abnormal classes; reported contamination ratios are 0%, 1%, 5%, 10%, and 20%.	Yes	Yes	Reports ROC-AUC under increasing contamination and shows improved or competitive AE robustness. The method is AE-specific and does not compare calibration with other mitigation mechanisms.
Adversarial AE rejection (Beggel et al., 2020)	Generic AD; image anomaly detection. Image outliers contaminate the training set; reported anomaly rates are 5%, 1%, and 0.1%, and latent-likelihood rejection filters suspected anomalies during adversarial AE training.	Yes	No	Reports balanced accuracy under image contamination and improved robustness over the adversarial AE baseline. It does not compare the rejection mechanism with other contamination-mitigation approaches.

3.1. Experiments

In Experiment 1, each dataset is partitioned into K folds. For datasets with multiple independent trajectories, folds are built asset-wise so that all samples from a given asset are assigned to a single fold. For each fold f, hyperparameters are selected on D_{-f}, and the selected configuration is then fitted transductively on all inputs. Metrics are computed only on the held-out fold D_f.

In Experiment 2, we reuse the best hyperparameters found in Experiment 1. For each fold and contamination ratio c, the model is trained on $D_{-f,c}$ and evaluated on the fixed held-out fold D_f. The contamination ratio is increased by progressively reducing the number of normal samples in the training split, while keeping the anomalous samples and the held-out fold unchanged.

We use *PyTorch* for model implementation (Paszke et al., 2019), and hyperparameter tuning is conducted with *Optuna* (Akiba, Sano, Yanase, Ohta, & Koyama, 2019). A fixed seed is used throughout the experiments, and the code is available

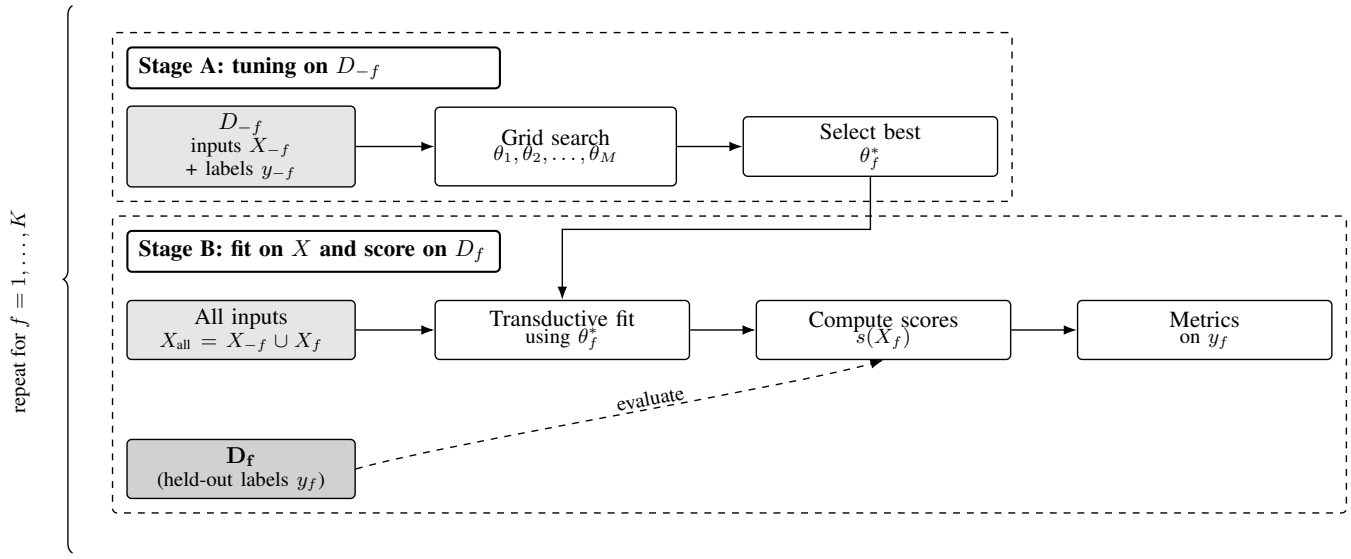

Figure 2. First comparison. Two-stage K-fold protocol: fold-specific hyperparameter selection on D_{-f} followed by transductive fitting on all inputs and evaluation on the held-out fold D_f only.

in a public GitLab repository.[1]

3.2. Data

The benchmark uses four public datasets selected to cover tabular and time-series data, as well as discrete anomalies and gradual degradation (Table 2). In all four cases, contamination is caused by anomalous or degraded samples being present in the training set.

3.2.1. Server Machine Dataset (SMD)

The Server Machine Dataset (SMD) (Su et al., 2019) is a multi-sensor time-series benchmark collected from 28 server machines. It contains abrupt annotated anomalies and is used here as a discrete time-series anomaly-detection task.

3.2.2. APS Failure at Scania Trucks

The APS dataset (Scania CV AB, 2016) is a high-dimensional tabular benchmark for fault diagnosis in heavy-duty trucks. The positive class indicates an Air Pressure System failure, while the negative class contains other failures and clean samples, resulting in a strongly imbalanced industrial classification problem.

3.2.3. Condition Based Maintenance of Naval Propulsion Plants (CBM)

The CBM dataset is a tabular degradation-monitoring benchmark for naval propulsion systems (Coraddu et al., 2014). It contains 11,934 samples with 16 features and two decay indicators, kMc and kMt; degraded samples are defined using the threshold ranges also used by Tan et al. (Tan et al., 2020).

3.2.4. FEMTO/PRONOSTIA Bearing Run-to-Failure

The FEMTO bearing dataset from the PRONOSTIA platform (Nectoux et al., 2012) contains run-to-failure vibration trajectories acquired under different operating speeds. Each vibration burst was transformed into 24 engineered time-, envelope-, and frequency-domain features computed from the horizontal and vertical accelerometers, with frequency-domain descriptors based on Welch power spectral density estimates (Welch, 1967).

3.3. Metrics

For SMD, Scania, and CBM, the task can be evaluated as one-class classification. Because anomalies are rare, we report AUPRC, best-F1 obtained by sweeping the decision threshold, and Recall@k over the highest-scoring alarms. For FEMTO, no direct degradation target is available, so we evaluate whether the anomaly score behaves like a degradation indicator.

Under the no-intervention assumption, a valid degradation indicator should be non-decreasing and should become largest near the end of the trajectory. We therefore use *Monotonicity* (Coble & Hines, 2009), originally introduced for prognostic-parameter evaluation:

$$\text{Monotonicity} = \frac{\sum_{j=1}^{n-1} |x_{j+1} - x_j| \ \text{sign}\big(x_{j+1} - x_j\big)}{\sum_{j=1}^{n-1} |x_{j+1} - x_j|}, \quad (1)$$

[1] https://gitlab.cylab.be/smac/mlpm/phme2026

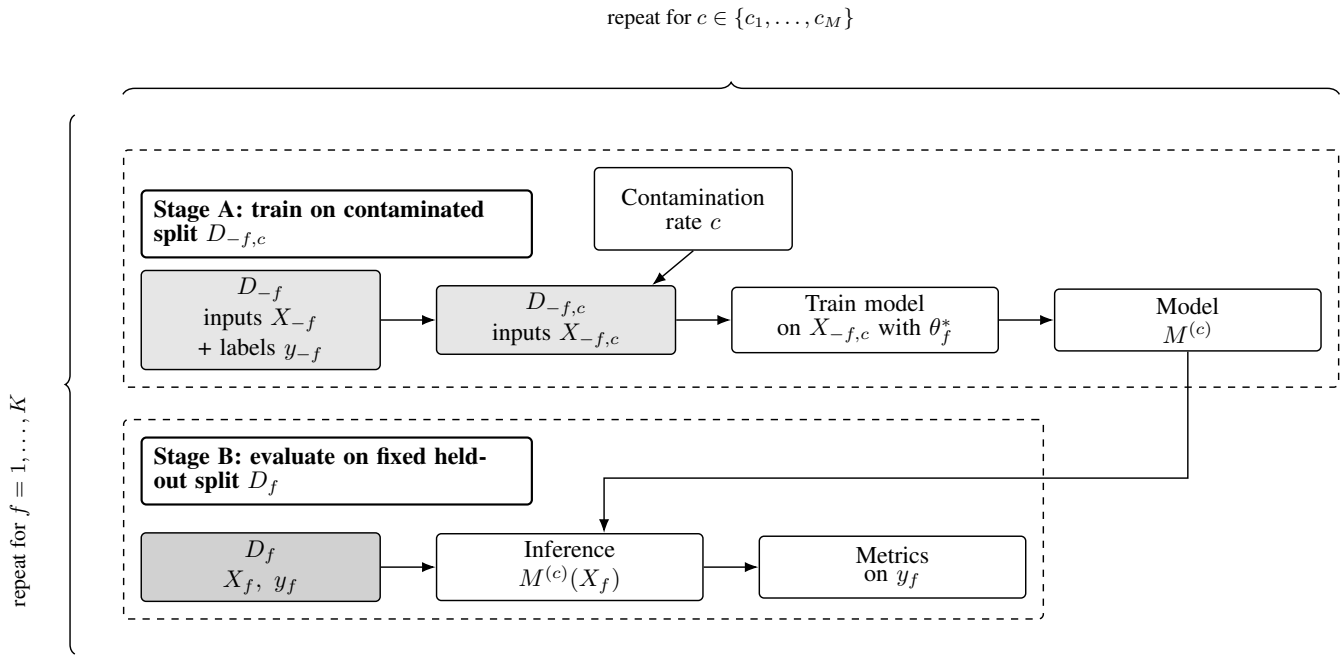

Figure 3. Second comparison. Inductive protocol: train on $D_{-f,c}$ with increasing contamination ratio c, evaluate on fixed held-out D_f, and report performance vs. c. In $D_{-f,c}$, the contamination ratio c is increased by downsampling normal samples in the training split.

Table 2. Main characteristics of the datasets used in this study.

Dataset	Nature	# Samples	# Features	# Trajectories	Anomaly / Target
SMD (OmniAD) (Su et al., 2019)	Multivariate time series	708000	38	28	Point-wise anomaly labels
Scania APS Failure (Scania CV AB, 2016)	Tabular	76000	170	1	Binary failure (APS)
CBM (Coraddu et al., 2014)	Tabular	11934	16	1	Continuous degradation (2 decay indicators)
FEMTO Bearing (Nectoux et al., 2012)	Time series (vibration)	87450[a]	24[a]	17	Continuous degradation (run-to-failure)

[a] Values obtained after pre-processing of the dataset.

Table 3. Compared methods in this benchmark.

Method	Family	Base model
AE	Reconstruction baseline	Dense AE
OCSVM	Classical OCC	n/a
Isolation Forest	Unsupervised outlier detection	n/a
WIRACAD	Iterative refinement	Dense AE
SRR	Iterative refinement	Dense AE
USDR	Subset refinement	Dense AE
RDA	Robust reconstruction	AE
READ-unsup.	Subset selection	RL-based
READ-informed	Limited supervision	RL-based

where x_j is the degradation indicator at sample j, and n is the number of samples in the trajectory. Finally, we report runtime as the mean computation time over folds for the selected hyperparameters.

3.4. Compared methods and benchmark instantiation

The AE baseline is implemented as a simple fully connected reconstruction model with one dense encoder layer and one dense decoder layer. Although more expressive variants, such as LSTM autoencoders (Srivastava, Mansimov, & Salakhudinov, 2015), can be better suited to time-series data, comparing AE architectures is outside the scope of this benchmark. WIRACAD, USDR, and SRR are therefore instantiated with the same dense AE backbone, so that performance differences mainly reflect the contamination-mitigation mechanism rather than the underlying architecture. This controlled setup compares refinement strategies directly; in practice, model-agnostic refinement methods could be coupled with more complex PHM backbones.

OCSVM (Schölkopf, Williamson, Smola, Shawe-Taylor, & Platt, 2000) is included as a classical one-class baseline. It does not explicitly correct for anomalous samples already present in the training set, making it a useful reference for contamination sensitivity. Isolation Forest (Liu, Ting, & Zhou, 2008) is included as a classical unsupervised outlier-detection baseline, providing a contrast to reconstruction-based methods because it isolates rare samples rather than learning an explicit normal reconstruction model.

WIRACAD, USDR, and SRR are refinement frameworks that

can be coupled to an underlying one-class detector. In this benchmark, all three are instantiated with the same AE backbone. WIRACAD (Donné et al., 2025) iteratively assigns continuous weights to training samples according to reconstruction residuals, reducing the influence of samples that remain difficult to reconstruct. USDR (Ulmer et al., 2024) trains models on multiple subsets and compares the residual behavior of samples when they are included in training versus when they are only inferred. SRR (Yoon et al., 2022) iteratively removes a fraction of the highest-scoring samples, retrains the one-class model, and uses the last refinement step for scoring.

READ (Shou et al., 2025) is included because it explicitly targets contamination and limited supervision through a reinforcemer learning subset-selection strategy. It builds on DQN-style learning (Mnih et al., 2013) and prior RL-based anomaly-detection work (Pang, van den Hengel, Shen, & Cao, 2021). We evaluate both the informed version, which uses labels for 5% of clean training data, and an unsupervised variant without label information.

RDA (Zhou & Paffenroth, 2017) is a contamination-aware reconstruction method inspired by robust PCA (Candès, Li, Ma, & Wright, 2009). It decomposes the input into a low-dimensional component reconstructed by the autoencoder and a sparse residual component that captures outliers and noise, making it a relevant baseline for contaminated AE training.

4. RESULTS

4.1. Transductive training across all datasets

The first comparison shows that contamination-mitigating methods can improve anomaly detection performance, but the gain is strongly dataset-dependent. The full results are reported in the Appendix, while Figure 4 summarizes them by reporting the arithmetic mean performance across datasets.

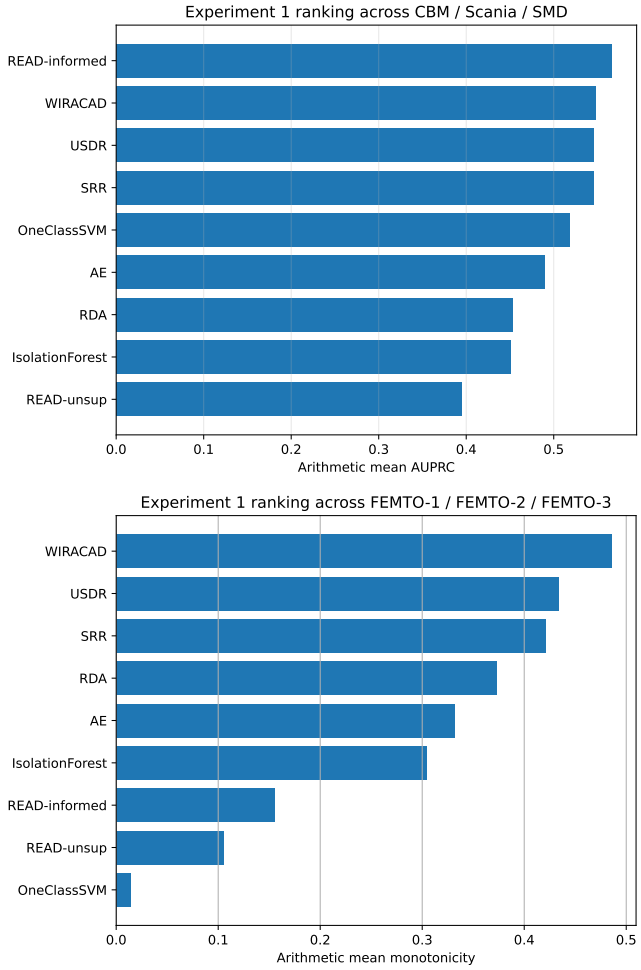

Figure 4. Aggregated performance of the evaluated methods in Experiment 1. Top: arithmetic mean AUPRC across CBM, Scania, and SMD. Bottom: arithmetic mean monotonicity across FEMTO-1, FEMTO-2, and FEMTO-3.

Overall, the results indicate that contamination-mitigating techniques can enhance the anomaly detection capabilities of semi-supervised methods when a purely normal training set cannot be isolated. This is especially true on the SMD and FEMTO datasets where the largest gap between refinement techniques and baselines is recorded. However, this remains a *no-free-lunch* setting, as no single method consistently outperforms the others across all benchmarks.

The ranking of methods varies substantially from one dataset to another. When aggregated across datasets (Figure 4), the refinement-based techniques (USDR, SRR, WIRACAD) exhibit very similar performance, making it difficult to draw a clear distinction between them. Among these methods, WIRACAD appears to perform slightly better on the FEMTO datasets. This observation is consistent with our previous work on the C-MAPSS dataset, where the method proved particularly effective at producing degradation indicators in settings involving continuous degradation processes (Donné et al., 2025).

We hypothesize that the continuous weighting mechanism used by WIRACAD during its iterative refinement may explain its stronger performance on datasets characterized by gradual degradation processes like in CBM or FEMTO. In contrast, SRR relies on a discrete removal strategy, where samples identified as anomalous are progressively excluded from the training set. This discrete refinement mechanism appears better suited to datasets such as SCANIA and SMD, where anomalies manifest as discrete events rather than progressive degradation patterns.

The informed version of READ also slightly outperformed the refinement-based techniques on the CBM, Scania, and SMD datasets, while the unsupervised version, which constitutes a fairer comparison, was the weakest method evaluated here. This result nevertheless demonstrates the benefit of

limited supervision by including even a small amount of prior information, since 5% of labeled clean data was sufficient to improve performance.

4.2. Inductive training with increasing contamination rate

The purpose of this comparison is to measure how different methods behave under increasing levels of contamination. The experiment is conducted on the SMD and Scania datasets, where anomalies are clearly defined as discrete binary events. We do not include FEMTO and CBM in this analysis because they exhibit continuous degradation processes, where samples can lie at intermediate degradation stages and the boundary between normal and abnormal states becomes less well defined. The contamination ratio was artificially increased by removing non-anomalous samples from the training set (as described in the Methodology section). The full results are reported in the Appendix.

On the Scania dataset, contamination mitigation techniques have a more robust response to increasing contamination ratio than the bare autoencoder. SRR in particular maintained a relatively high AUPRC even for high contamination ratio (46%). IForest, as an unsupervised outlier detection technique, remained comparatively insensitive to increasing contamination in this setting. Figure 5 illustrates these results for four selected techniques.

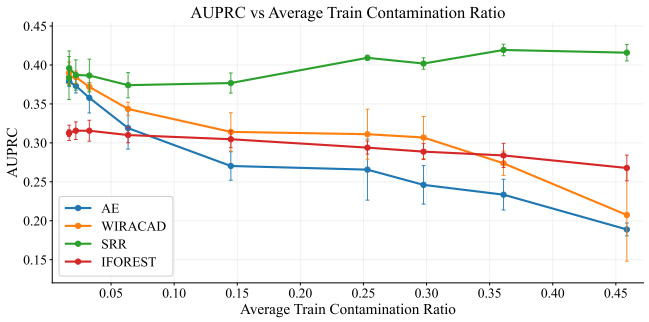

Figure 5. Evolution of the evaluated AUPRC for 4 different methods with increasing contamination ratio on the SCANIA dataset. IForest is the baseline and the AE the standard reconstruction technique subjected to contamination. Contamination mitigating techniques, WIRACAD and especially SRR here, exhibit a more robust response to increasing contamination.

This experiment also highlights one of the main effects of contamination on AD techniques that try to learn a normal representation: the line between anomalous and non-anomalous samples becomes blurry. This is illustrated in Figure 6 where we observe the residual distribution for two methods, at initial contamination ratio (1.7%) and at an advanced ratio (45.9%). For both methods, the separability between normal and anomalous samples is substantially reduced by increased contamination ratio. This effect is especially visible for READ, for which the distributions strongly overlap at the highest contamination ratio. READ was the worst performing method on the Scania dataset.

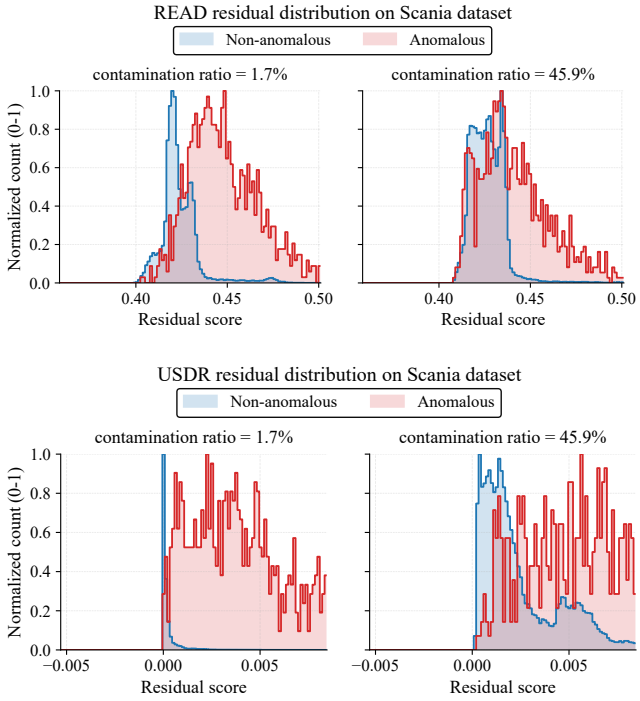

Figure 6. READ unsupervised (top) and USDR (bottom) residual distributions on Scania for two different contamination ratios. The separability between anomalous and non-anomalous samples is significantly lower for the model trained on the most contaminated dataset. READ unsupervised was the worst performing method on this dataset while USDR showed results similar to SRR.

Results differ sharply between Scania and SMD. On Scania, methods behave as expected: contamination-mitigation techniques are more robust as contamination increases. On SMD, increasing contamination has a much smaller impact (on average), with strong machine-to-machine variability; on some assets, AUPRC changes little even at higher contamination levels. An attempt to explain this variability is provided in 5.1.

5. DISCUSSION

5.1. Nature of the anomalies

In the second comparison, the effect of increasing contamination appears to have an important machine-to-machine variability on the SMD dataset. To better understand these results, we analyze the anomaly distribution with the Normalized Clusteredness (NC) (Emmott, Das, Dietterich, Fern, & Wong, 2013). This is fundamentally a ratio between the average pairwise distance of normal samples and average pairwise distance of anomalous samples. Let S_{nor} and S_{ano} denote the

sets of normal and anomalous samples, and $D(\cdot, \cdot)$ a distance in feature space. NC is defined as

$$\mathrm{NC} = \frac{\mathbb{E}_{x_i \neq x_j \in S_{\mathrm{nor}}}[D(x_i, x_j)]}{\mathbb{E}_{x_i \neq x_j \in S_{\mathrm{ano}}}[D(x_i, x_j)]}. \tag{2}$$

For each SMD machine, we measure the slope of the AUPRC decreasing curve with increasing contamination ratio. We focus on the autoencoder, used here as the representative contaminati sensitive method. The more negative the slope, the more important the AE performance degradation with increasing contamination.

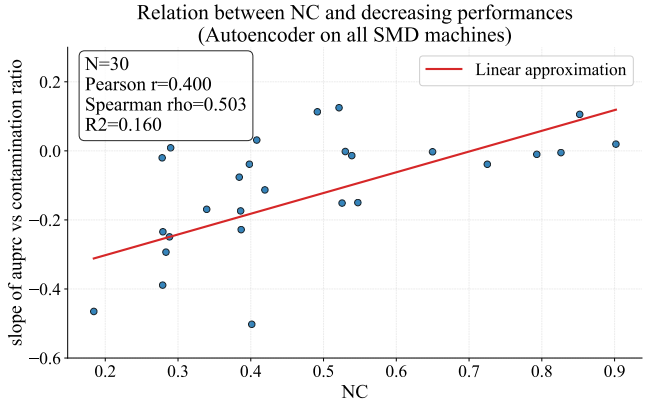

Figure 7. Relation between anomaly clusteredness and sensitivity to contamination on the SMD dataset. The y-axis represents the slope of the AUPRC variation with respect to contamination ratio: more negative values indicate a stronger degradation of performance as contamination increases.

We observe a positive association between NC and the AE robustness to contamination: lower NC (i.e., anomalies more spread across the dataset) tends to correspond to a steeper AUPRC drop as contamination increases. Under the standard two-sided Pearson correlation test ($H_0 : \rho = 0$), the observed correlation $r = 0.400$ with $n = 30$ yields $t = r\sqrt{(n-2)/(1-r^2)} = 2.309$, corresponding to a p-value $p = 0.0285 < 5\%$. While this association is statistically significant, it is not strong enough to draw a definitive conclusion about the phenomenon we observed on the SMD dataset. It is however sufficient to suggest that the distribution of the anomalies in the dataset (their nature) is possibly impacting how AD methods react to contamination and that the contamination ratio itself is not sufficient to predict performance degradation. Scania has NC ≈ 0.12, which would place it in the low-NC regime of Fig. 7, consistent with the stronger contamination effects observed on this dataset.

Our observations on the SMD dataset suggest that the degradation induced by contaminated training data cannot be explained by the contamination ratio alone. In particular, performance appeared to correlate more strongly with the dispersion of anomalies, as reflected by the NC metric, than with the proportion of anomalous samples present in the training set. This suggests that some structural properties of anoma-

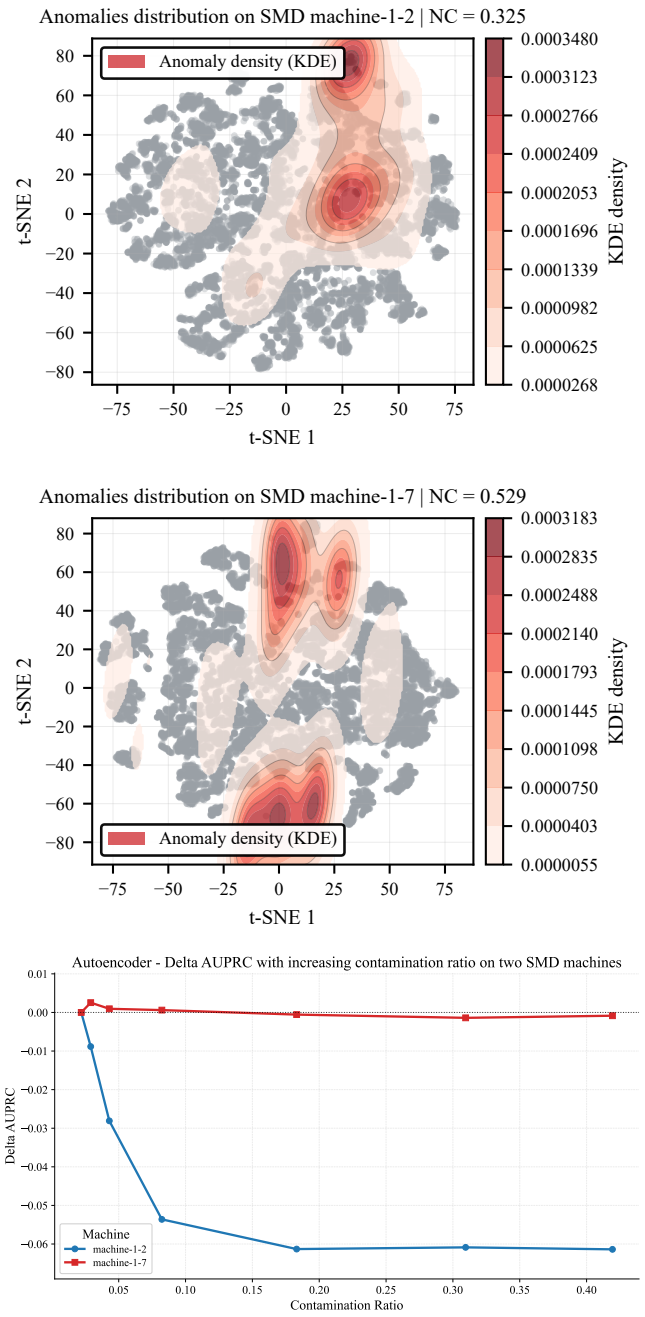

Figure 8. Illustration of the various anomalies distribution encountered in the SMD datasets. Here the machine-1-7 (middle) displays a relatively high NC which indicated that anomalies are relatively close to each other. Anomalies are more spread across the dataset for the machine-1-2 (top). The curves (bottom) show the difference of AUPRC from the initial values, with an increasing contamination ratio, for these two illustrated SMD machines.

lies may affect how strongly contamination alters the learning process. This behavior is illustrated in Fig. 8. For machine-1-7, which exhibits a relatively high NC, the AE performance remains almost unchanged as the contamination ratio increases. Conversely, machine-1-2 displays a lower NC with more dispersed anomalies, and the AE performance rapidly degrades as contamination increases. One possible interpretation is that dispersed anomalies may affect the learned representation of normality more broadly than tightly clustered anomalies, whose effect may remain more localized. However, this interpretation should be taken with caution. NC only measures how close anomalies are to one another; it does not capture how close they are to the support of normal data. As a result, NC should be viewed here as an informative descriptor of anomaly structure, rather than a complete explanation of contamination sensitivity.

The anomaly detection literature has previously shown that the nature of anomalies significantly influences detection difficulty. For instance, benchmark studies have highlighted that anomalies that lie close to the support of normal data or that form compact clusters can affect the behavior of anomaly detection algorithms in different ways (Emmott et al., 2013). More recently, large-scale evaluations have further confirmed that algorithm performance varies substantially depending on anomaly characteristics and structure (Han, Hu, Huang, Jiang, & Zhao, 2022). While these observations were established in the context of anomaly detection tasks rather than contaminated training settings, they suggest a plausible mechanism where the properties of anomalous samples may influence how a one-class model trained on contaminated data learns a representation of the normal state. Under this hypothesis, dispersed anomalies or anomalies located near the normal data manifold may more easily influence the learned representation of the normal state, whereas tightly clustered anomalies may have a more localized impact on the learned distribution. Future research should aim to bridge findings established in classical AD settings with contaminated AD, in order to better characterize how the properties of anomalies influence the learning of the normal representation of a system.

5.2. PHM-specific contamination sources

In the literature, and in the context of this work, contamination typically refers to the presence of abnormal points in the training set of a semi-supervised method that aims to learn a representation of the normal behavior of a system and extrapolate it to perform anomaly detection (Qiu et al., 2022; Shou et al., 2025). However, in real PHM applications, several additional phenomena can hinder the learning of a clean representation of normal operation beyond the presence of anomalies. For instance, operating-condition variability can further complicate the learning of a normal representation, as the definition of a normal state may vary significantly across different operating conditions or system loads (Zhao et al., 2019). Detecting transitions between operating conditions is important to create a proper representation of a normal state (Kermenov et al., 2023). Failing to do so may lead to incorporating information from operating conditions *B* or *C* into the learned representation of condition *A*. Future research should aim to study these phenomena within the framework of contamination, using similar tools, in order to better understand how strongly they can impact AD capabilities.

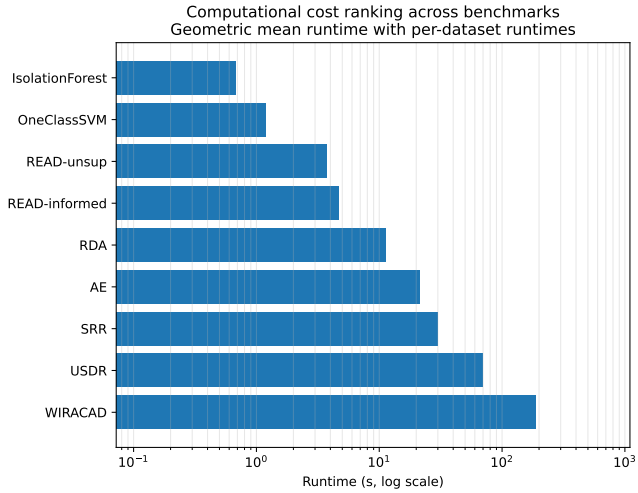

Figure 9. Computational-cost ranking of the evaluated methods (Experiment 1) Tables 4 to 7. Bars show the geometric mean runtime (s) across datasets and the x-axis is logarithmic.

5.3. Computational cost

The computational time of the different frameworks is compared using the results reported in the tables of the first experiment. All experiments were executed on a laptop running on Windows 11 and equipped with an Intel i7-12850HX processor, 32 GB of RAM, and an NVIDIA RTX A3000 GPU with 12 GB of memory.

To summarize computational cost across datasets with widely different runtime scales, we report the geometric mean of runtimes, computed as

$$GM = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\ln(t_i)\right) \tag{3}$$

where t_i denotes the runtime measured on dataset i and n is the number of datasets considered.

In our setting, classical methods such as Isolation Forest run almost instantaneously, whereas deep-learning based techniques are significantly slower. In particular, iterative refinement frameworks (SRR, USDR, WIRACAD) are the most computationally expensive, as they require training many AutoEncoders (AEs) across multiple iterations and refinement steps.

The runtime strongly depends on the hyperparameters selected during tuning. For instance, WIRACAD with $n\_fold = 3$ and $n\_iteration = 4$ requires training 12 AEs. By contrast, $n\_fold = 5$ and $n\_iteration = 10$ results in 50 AE trainings.

These results indicate that the benefits of contamination mitigation should be considered in light of the computational cost. Iterative frameworks, in particular, may be impractical for time-sensitive applications, for streaming data processing or resource-constrained embedded computing.

6. CONCLUSION

This work investigated the impact of contaminated training data on anomaly detection methods used for diagnostics in PHM systems. Because semi-supervised anomaly detection typically assumes access to clean normal data, even small amounts of anomalous samples in the training set can degrade the learned representation of normal behavior. To assess how different approaches handle this issue, we evaluated several contamination-aware techniques alongside classical baselines across four datasets representing diverse PHM contexts, including tabular and time-series data as well as discrete anomalies and gradual degradation scenarios.

The first comparison showed that contamination-mitigating techniques can enhance the anomaly-detection capabilities of semi-supervised methods when a purely normal training set cannot be isolated, although their effectiveness varies substantially across datasets. Under a unified benchmark, our results show strong dataset dependence, with no method emerging as a consistently reliable choice across the PHM settings considered here. This supports a no-free-lunch view of contaminated anomaly detection. The largest gains over classical baselines were observed on the SMD and FEMTO datasets, suggesting that refinement-based approaches may be particularly beneficial in the time-series settings considered here.

Among refinement-based methods, performance was broadly similar but with noticeable dataset-dependent differences: WIRACAD tended to perform better on datasets characterized by gradual degradation processes, whereas SRR obtained stronger results on datasets with discrete anomalies. This contrast is consistent with their respective refinement mechanisms, namely continuous sample weighting for WIRACAD and iterative discrete sample removal for SRR. While these trends are observed on only four datasets and should not be overgeneralized, they suggest that the effectiveness of contamination-mitigating methods depends on the interaction between method design and anomaly structure, rather than on contamination ratio alone.

In the second comparison, refinement techniques were globally more robust to increasing contamination than sensitive baselines such as a plain autoencoder. This trend was particularly clear on Scania, where methods such as SRR maintained stronger performance as contamination increased, while on SMD the effect of contamination was smaller on average and more heterogeneous across machines. More broadly, these results suggest that the degradation in performance cannot be explained solely by the contamination ratio; the structure and distribution of anomalies in the data also appear to influence how strongly models are affected.

Future work should therefore investigate the role of anomaly structure and extend the notion of contamination to PHM-specific phenomena that may hinder the learning of a normal representation of a system. In particular, it should (i) determine whether PHM-specific contamination mechanisms, such as early degradation stages or operating regime changes, produce distinct anomaly structures, and (ii) evaluate a wider range of contamination-mitigation methods across these different structures to identify the conditions under which each method is most effective.

REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining.* doi: 10.1145/3292500.3330701

Beggel, L., Pfeiffer, M., & Bischl, B. (2020). Robust anomaly detection in images using adversarial autoencoders. In *Machine learning and knowledge discovery in databases* (Vol. 11906, pp. 206–222). Springer. doi: 10.1007/978-3-030-46150-8_13

Candès, E. J., Li, X., Ma, Y., & Wright, J. (2009). *Robust principal component analysis?* arXiv preprint arXiv:0912.3599. doi: 10.48550/arXiv.0912.3599

Carvalho, T. P., Soares, F. A., Vita, R., Francisco, R. d. P., Basto, J. P., & Alcala, S. G. S. (2022). Artificial intelligence and real-time predictive maintenance in Industry 4.0: A bibliometric analysis. *Artificial Intelligence Review*, *55*, 2127–2170. doi: 10.1007/s10462-021-10082-7

Chandola, V., Banerjee, A., & Kumar, V. (2009, July). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), 15:1–15:58. doi: 10.1145/1541880.1541882

Coble, J., & Hines, J. W. (2009). Identifying optimal prognostic parameters from data: A genetic algorithms approach. In *Proceedings of the annual conference of the prognostics and health management society* (Vol. 1). (Available from the PHM Society proceedings archive)

Coraddu, A., Oneto, L., Ghio, A., Savio, S., Anguita, D., & Figari, M. (2014, September). *Condition based maintenance of naval propulsion plants [dataset].* UCI Machine Learning Repository. doi: 10.24432/C5K31K

del Moral, P., Nowaczyk, S., & Pashami, S. (2022). Filtering misleading repair log labels to improve predictive maintenance models. In *Proceedings of the 7th European Conference of the Prognostics and Health Management Society 2022* (Vol. 7, pp. 110–117). Prognostics and Health Management Society. doi: 10.36001/phme.2022.v7i1.3360

Donné, S., Davis, J., Van Utterbeeck, F., & Verbeke, M. (2025, October). Anomaly detection under contaminated data: A weighted iterative refinement framework for health monitoring. In *2025 IEEE 12th international conference on data science and advanced analytics.* Birmingham, United Kingdom: IEEE. doi: 10.1109/DSAA65442.2025.11248011

Emmott, A. F., Das, S., Dietterich, T., Fern, A., & Wong, W.-K. (2013). Systematic construction of anomaly detection benchmarks from real data. In *Proceedings of the ACM SIGKDD workshop on outlier detection and description* (pp. 16–21). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2500853.2500858

Han, S., Hu, X., Huang, H., Jiang, M., & Zhao, Y. (2022). ADBench: Anomaly detection benchmark. *Advances in Neural Information Processing Systems*, *35*, 32142–32159.

Huang, Y., Tang, Y., VanZwieten, J. H., & Liu, J. (2020). Reliable machine prognostic health management in the presence of missing data. *Concurrency and Computation: Practice and Experience*, *34*(12), e5762. doi: 10.1002/cpe.5762

Kermenov, R., Nabissi, G., Longhi, S., & Bonci, A. (2023). Anomaly detection and concept drift adaptation for dynamic systems: A general method with practical implementation using an industrial collaborative robot. *Sensors*, *23*(6), 3260. doi: 10.3390/s23063260

Kim, M., Yu, J., Kim, J., Oh, T.-H., & Choi, J. K. (2024, October). An iterative method for unsupervised robust anomaly detection under data contamination. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(10), 13327–13339. doi: 10.1109/TNNLS.2023.3267028

Li, Y., Kang, J.-M., & Kim, I.-M. (2025, June). Beyond clean training data: A versatile and model-agnostic framework for out-of-distribution detection with contaminated training data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10183–10192). IEEE. doi: 10.1109/CVPR52734.2025.00952

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *2008 eighth IEEE international conference on data mining* (pp. 413–422). IEEE. doi: 10.1109/ICDM.2008.17

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). *Playing Atari with deep reinforcement learning.* arXiv preprint arXiv:1312.5602. doi: 10.48550/arXiv.1312.5602

Munir, M., Siddiqui, S. A., Dengel, A., & Ahmed, S. (2019). DeepAnT: A deep learning approach for unsupervised anomaly detection in time series. *IEEE Access*, *7*, 1991–2005. doi: 10.1109/ACCESS.2018.2886457

Nectoux, P., Gouriveau, R., Medjaher, K., Ramasso, E., Morello, B., Zerhouni, N., & Varnier, C. (2012, June). PRONOSTIA: An experimental platform for bearings accelerated degradation tests. In *Proceedings of the IEEE international conference on prognostics and health management* (pp. 1–8). Denver, CO, USA: IEEE. doi: 10.1109/PHM.2012.6228857

Pang, G., van den Hengel, A., Shen, C., & Cao, L. (2021). Toward deep supervised anomaly detection: Reinforcement learning from partially labeled anomaly data. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 1298–1308). Association for Computing Machinery. doi: 10.1145/3447548.3467417

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... others (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems* (Vol. 32). Curran Associates.

Qiu, C., Li, A., Kloft, M., Rudolph, M., & Mandt, S. (2022). Latent outlier exposure for anomaly detection with contaminated data. In *Proceedings of the 39th international conference on machine learning* (Vol. 162, pp. 18153–18167). PMLR.

Scania CV AB. (2016, September). *APS failure at Scania trucks [dataset].* UCI Machine Learning Repository. doi: 10.24432/C51S51

Schmidl, S., Wenig, P., & Papenbrock, T. (2022). Anomaly detection in time series: A comprehensive evaluation. *Proceedings of the VLDB Endowment*, *15*(9), 1779–1797. doi: 10.14778/3538598.3538602

Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., & Platt, J. C. (2000). Support vector method for novelty detection. In *Advances in neural information processing systems 12* (pp. 582–588).

Shou, H., Lu, G., Pavlovski, M., & Zhou, F. (2025, August). READ: Robust and efficient anomaly detection under data contamination and limited supervision. In *Proceedings of the 31st ACM SIGKDD conference on knowledge discovery and data mining* (pp. 2586–2596). Toronto, ON, Canada: Association for Computing Machinery. doi: 10.1145/3711896.3737100

Srivastava, N., Mansimov, E., & Salakhudinov, R. (2015). Unsupervised learning of video representations using LSTMs. In *Proceedings of the 32nd international conference on machine learning* (Vol. 37, pp. 843–852). Lille, France: PMLR.

Su, Y., Liu, R., Zhao, Y., Sun, W., Niu, C., & Pei, D. (2019, August). Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2828–2837). Anchorage, AK, USA: Association for Computing Machinery. doi: 10.1145/3292500.3330672

Tan, Y., Tian, H., Jiang, R., Lin, Y., & Zhang, J. (2020, April). A comparative investigation of data-driven approaches based on one-class classifiers for condition monitoring of marine machinery system. *Ocean Engineering*, *201*, 107174. doi: 10.1016/j.oceaneng.2020.107174

Ulmer, M., Zgraggen, J., & Goren Huber, L. (2024, January). A generic machine learning framework for fully-unsupervised anomaly detection with contaminated data. *International Journal of Prognostics and Health Management*, *15*(1), 1–12. doi: 10.36001/ijphm.2024.v15i1.3589

Wagner, D., Michels, T., Schulz, F. C. F., Nair, A., Rudolph, M., & Kloft, M. (2023). TimeSeAD: Benchmarking deep multivariate time-series anomaly detection. *Transactions on Machine Learning Research*. (Available at OpenReview)

Welch, P. D. (1967, June). The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, *15*(2), 70–73. doi: 10.1109/TAU.1967.1161901

Wen, J., Ren, J., Zhao, Z., & Chen, X. (2025, August). Robust anomaly detection of rotating machinery with contaminated data. *Journal of Dynamics, Monitoring and Diagnostics*, *4*(3), 170–182. doi: 10.37965/jdmd.2025.855

Yoon, J., Sohn, K., Li, C.-L., Arik, S. O., Lee, C.-Y., & Pfister, T. (2022). Self-supervise, refine, repeat: Improving unsupervised anomaly detection. *Transactions on Machine Learning Research*. (Published August 4, 2022)

Yu, J., Kim, M., Kim, J., & Oh, H. (2026). Normality-calibrated autoencoder for unsupervised anomaly detection on data contamination. *Neurocomputing*, *667*, 132249. doi: 10.1016/j.neucom.2025.132249

Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*, *115*, 213–237. doi: 10.1016/j.ymssp.2018.05.050

Zhou, C., & Paffenroth, R. C. (2017). Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 665–674). Halifax, NS, Canada: Association for Computing Machinery. doi: 10.1145/3097983.3098052

APPENDIX

Table 4. CBM benchmark metrics (mean ± std across folds) - First comparison.

Method	Best-F1	AUPRC	Rec@k	Runtime (s)
Baselines				
Plain AE	0.85 ±0.0006	0.77 ±0.0129	0.73 ±0.0174	4.09 ±2.313
IsolationForest	0.74 ±0.0052	0.57 ±0.0174	0.56 ±0.0097	0.33 ±0.0060
OneClassSVM	0.86 ±0.0011	0.76 ±0.0101	0.76 ±0.0065	0.15 ±0.0644
Contamination-robust / refinement				
WIRACAD	0.85 ±0.0001	0.77 ±0.0103	0.76 ±0.0034	163.22 ± 134.4070
RDA	0.85 ±0.0001	0.78 ±0.0044	0.76 ±0.0102	7.24 ±0.4078
SRR	0.74 ±0.0105	0.62 ±0.0313	0.60 ±0.0311	6.72 ±3.411
USDR	0.85 ±0.0008	0.83 ±0.0068	0.75 ±0.0061	20.43 ± 6.025
READ-unsup	0.85 ±0.0006	0.78 ±0.0052	0.74 ±0.0069	3.31 ±0.5905
READ-informed	0.86 ±0.0064	0.77 ±0.0096	0.73 ±0.0095	5.13 ±2.4554

Table 5. FEMTO benchmark metrics (mean ± std across assets) - First comparison. Monotonicity is reported separately for FEMTO-1/2/3, while runtime is aggregated across the three FEMTO dataset.

Method	Monotonicity (FEMTO-1)	Monotonicity (FEMTO-2)	Monotonicity (FEMTO-3)	Runtime (s)
Baselines				
AE	0.33 ±0.3583	0.31 ±0.4439	0.35 ±0.4130	6.17 ±2.7686
IsolationForest	0.28 ±0.2894	0.19 ±0.1760	0.44 ±0.2672	0.35 ±0.0356
OneClassSVM	0.11 ±0.1010	-0.17 ±0.0961	0.11 ±0.5289	0.34 ±0.2988
Contamination-robust / refinement				
WIRACAD	0.51 ±0.4561	0.24 ±0.1836	0.70 ±0.1502	112.50 ±31.4666
RDA	0.41 ±0.4855	0.11 ±0.1322	0.60 ±0.5039	6.84 ±6.4500
SRR	0.40 ±0.3692	0.26 ±0.3331	0.60 ±0.1824	22.06 ±8.4887
USDR	0.42 ±0.3395	0.20 ±0.2008	0.68 ±0.1673	24.81 ±9.1067
READ-unsup	0.14 ±0.0856	0.00 ±0.0264	0.17 ±0.2843	2.83 ±0.3958
READ-informed	0.07 ±0.0668	0.24 ±0.1413	0.15 ±0.2113	3.00 ±0.3903

Table 6. SCANIA benchmark metrics (mean ± std across folds) - First comparison.

Method	Best-F1	AUPRC	Rec@k	Runtime (s)
Baselines				
Plain AE	0.50 ±0.0202	0.41 ±0.0334	0.46 ±0.0315	102.63 ±55.6284
IsolationForest	0.42 ±0.0114	0.37 ±0.0028	0.39 ±0.0107	8.72 ±7.7543
OneClassSVM	0.47 ±0.0374	0.37 ±0.0408	0.45 ±0.0472	98.90 ±0.5725
Contamination-robust / refinement				
WIRACAD	0.48 ±0.0133	0.40 ±0.0444	0.46 ±0.0184	333.33 ±97.1467
RDA	0.48 ±0.0284	0.40 ±0.0467	0.44 ±0.0448	80.97 ±1.7880
SRR	0.54 ±0.0013	0.51 ±0.0326	0.52 ±0.0141	87.23 ±25.0538
USDR	0.35 ±0.0077	0.18 ±0.0025	0.34 ±0.0099	74.41 ±5.4045
READ-unsup	0.19 ±0.0068	0.07 ±0.0009	0.00 ±0.0014	5.35 ±1.3175
READ-informed	0.60 ±0.0190	0.62 ±0.0451	0.59 ±0.0226	6.70 ±1.1871

Table 7. SMD benchmark metrics (mean ± std across assets) - First comparison.

Method	Best-F1	AUPRC	Rec@k	Runtime (s)
Baselines				
AE	0.51 ±0.1787	0.46 ±0.2064	0.46 ±0.2042	41.65 ±4.1545
IsolationForest	0.47 ±0.1765	0.42 ±0.1968	0.42 ±0.1671	0.74 ±0.0473
OneClassSVM	0.48 ±0.1989	0.42 ±0.2046	0.45 ±0.1897	15.56 ±4.2552
Contamination-robust / refinement				
WIRACAD	0.52 ±0.2045	0.47 ±0.1972	0.46 ±0.1876	606.91 ±59.4080
RDA	0.26 ±0.1559	0.18 ±0.1139	0.22 ±0.1025	40.15 ±12.2357
SRR	0.55 ±0.2204	0.51 ±0.2387	0.51 ±0.2242	155.82 ±13.3897
USDR	0.52 ±0.2162	0.46 ±0.2517	0.43 ±0.2578	171.70 ±28.8253
READ-unsup	0.40 ±0.2047	0.33 ±0.2060	0.35 ±0.2130	6.96 ±0.5549
READ-informed	0.39 ±0.1803	0.30 ±0.1833	0.33 ±0.1792	11.09 ±3.8089

Table 8. SMD: relative AUPRC change (%) with respect to the baseline model trained with 4.5% contamination. Second comparison.

Method	Training contamination ratio						
	5.8%	**8.3%**	**14.8%**	**28.4%**	**42.0%**	**49.5%**	**61.2%**
AE	-5.8%	-16.0%	-23.5%	-26.9%	-28.8%	-29.3%	-29.3%
IForest	-1.4%	-6.2%	-14.9%	-32.8%	-41.6%	-38.8%	-49.5%
OCSVM	-3.1%	-8.6%	-21.1%	-37.7%	-45.8%	-48.1%	-47.4%
WIRACAD	-6.0%	-11.9%	-22.5%	-26.4%	-27.2%	-26.6%	-26.8%
RDA	-3.0%	-11.4%	-23.4%	-31.8%	-36.1%	-61.6%	-43.9%
SRR	-2.8%	-9.4%	-14.2%	-22.9%	-23.1%	-6.2%	-24.5%
USDR	-8.6%	-17.6%	-23.7%	-23.4%	-23.1%	-23.1%	-23.1%
READ-unsup.	-0.4%	-2.6%	-9.3%	-20.0%	-20.1%	-19.3%	-32.4%
READ-informed	12.8%	-1.0%	-18.3%	-19.9%	-36.1%	-44.4%	-50.2%

Table 9. SCANIA: relative AUPRC change (%) with respect to the baseline model trained with 1.67% contamination. Second comparison.

Method	Training contamination ratio					
	2.21%	**3.28%**	**6.35%**	**14.5%**	**25.3%**	**45.9%**
AE	-2.7%	-5.5%	-21.0%	-28.6%	-29.8%	-55.4%
IForest	-1.0%	+1.0%	-0.8%	-2.5%	-5.9%	-14.3%
OCSVM	-1.4%	-3.5%	-12.4%	-34.6%	-52.0%	-69.1%
WIRACAD	-1.4%	-4.6%	-11.9%	-19.4%	-20.2%	-46.8%
RDA	1.6%	7.6%	-7.2%	-43.4%	-59.7%	-87.8%
SRR	+1.0%	+0.8%	-2.4%	-1.7%	+6.8%	+8.5%
USDR	-1.3%	-2.8%	-9.4%	-13.8%	-15.9%	-24.9%
READ-unsup.	-2.7%	-6.6%	-8.1%	-1.6%	-26.0%	-39.7%
READ-informed	5.9%	-2.6%	-3.6%	-22.8%	-67.4%	-51.1%

Note. For both Tables 8 and 9, the lowest contamination level corresponds to the native anomaly proportion available in the training folds. Because contamination is increased by downsampling normal samples, lower ratios such as 0.1% would require either discarding anomalous samples or changing the dataset distribution, and were therefore not included.

Table 10. Benchmark tuning space (contamination-mitigating techniques only). 40 trials.

Method	Parameter	Range
SRR	`anomaly_ratio_percent`	$[2.0, 20.0]$
	`n_iterations`	$\{2, \ldots, 6\}$
	`ensemble_count`	$\{1, \ldots, 5\}$
USDR	`subset_fraction`	$[0.10, 0.50]$
	`n_subsets`	$\{5, \ldots, 50\}$
	`equal_representation`	$\{\texttt{false}, \texttt{true}\}$
WIRACAD	`weight_learning_rate`	$[5 \times 10^{-3}, 5 \times 10^{-1}]$ (log)
	`n_fold`	$\{3, \ldots, 6\}$
	`n_iterations`	$\{3, \ldots, 10\}$
READ-unsup	`contamination_rate`	$[0.01, 0.20]$
	`sampling_du`	$\{128, 256, \ldots, 1024\}$
READ-informed	`contamination_rate`	$[0.01, 0.20]$
	`sampling_du`	$\{128, 256, \ldots, 1024\}$
RDA	`lambda_per_sample`	$[10^{-5}, 3 \times 10^{-3}]$ (log)
	`inner_iteration`	$\{10, 15, \ldots, 40\}$
	`iteration`	$\{2, \ldots, 8\}$